INFORMATION AND CONTROL 2, 393–395 (1959)

A Note on Phrase Structure Grammars

NOAM CHOMSKY

Massachusetts Institute of Technology, Cambridge, Massachusetts

In Chomsky (1959)[1] a class of grammars is studied each of which contains a finite number of "rules" of the form $A \rightarrow \varphi$, where A is a single symbol and φ is not null. Such grammars (there called type 2 grammars) we will now call *context-free* (CF) phrase structure grammars. A sequence $(\varphi_1, \cdots, \varphi_n)$ of strings is called a *φ-derivation* of the CF grammar G if $\varphi = \varphi_1$ and for each $i < n$, there are strings $A, \psi_1, \psi_2, \psi_3$ such that $\varphi_i = \psi_1 A \psi_2$, $\varphi_{i+1} = \psi_1 \psi_3 \psi_2$ and $A \rightarrow \psi_3$ is a rule of G. φ is a *terminal string* if it contains no A such that for some ψ, $A \rightarrow \psi$ is a rule of G. The language L_G *generated by* G is the set of terminal strings that appear in (and thus conclude) S-derivations of G, where S is a designated initial symbol. A string is *derivable* if it is a step in an S-derivation. G is a *self-embedding* (s.e.) grammar if it contains strings A, φ_1, φ_2 such that φ_1 and φ_2 are not null and there is an A-derivation of $\varphi_1 A \varphi_2$. $\varphi \Rightarrow \psi$ if and only if ψ is a line of a φ-derivation.

Given a finite state Markov source Σ with a designated initial state S_0 and a symbol emitted with each interstate transition, we define the *language generated by* Σ as the set of strings produced as the system moves from S_0 to a first recurrence of S_0. The set of languages that can be generated in this way we call *finite state languages*.[2] Clearly finite state languages constitute a proper subset of the languages that can be generated by CF grammars (cf. Chomsky (1959), §5). It is an interesting and important problem to characterize precisely the set of nonfinite state languages that can be generated by CF grammars. As a step towards this, it was proven in Chomsky (1959) that a set of strings is

[1] The notations and terminology of that paper will be used in this note. In particular, we use the following notational convention: capital letters will be used for nonterminal strings (see below); small Latin letters for terminal strings; Greek letters for arbitrary strings; early letters of all alphabets for single symbols; late letters for arbitrary strings.

[2] Finite state languages are what are called "regular events" in Kleene (1956).

not a finite state language just in case all of its CF grammars are s.e. The basic theorem underlying this is the following.

THEOREM 1. If G is a non-s.e. CF grammar, then there is a finite state Markov source that generates the language L_G generated by G.

A method for constructing the equivalent finite state source, and a long and cumbersome proof of equivalence, was presented in Chomsky (1959). The purpose of this note is to present a much shorter and simpler proof of Theorem 1.

LEMMA 1. If G is a CF grammar generating L_G and every nonterminal derivable string is of the form xA or every nonterminal derivable string is of the form Ax, then there is a finite state Markov source that generates L_G.

LEMMA 2. Suppose that L_1 and L_2 are finite state languages and that a is a symbol in L_1. Let L_3 consist of all strings of L_1 that do not contain a and all strings formed by substituting a string of L_2 for each a occurring in a string of L_1. Then L_3 is a finite state language.

LEMMA 3. If L_1 and L_2 are finite state languages, then so are L_3 and L_4, where

(1) L_3 is the set of strings xy such that $x \in L_1$ and $y \in L_2$

(2) L_4 is the Boolean sum of L_1 and L_2.

Proof of Lemmas 1 and 2 is straightforward. For Lemma 3, see Kleene (1956).

Suppose now that G is a non-s.e. CF grammar generating the terminal language L.

(I) Suppose that G contains n symbols $A_1, \cdots, A_n$ and for each pair (i, j), there are strings φ, ψ such that $A_i \Rightarrow \varphi A_j \psi$. Suppose that for some i, j, k, l, $A_i \Rightarrow \varphi_1 A_j \varphi_2$ and $A_k \Rightarrow \psi_1 A_l \psi_2$, where φ_1 and ψ_2 are nonnull. Therefore $A_i \Rightarrow \varphi_1 A_j \varphi_2 \Rightarrow \varphi_1 \omega_1 A_k \omega_2 \varphi_2 \Rightarrow \varphi_1 \omega_1 \psi_1 A_l \psi_2 \omega_2 \varphi_2 \Rightarrow \varphi_1 \omega_1 \psi_1 \omega_3 A_i \omega_4 \psi_2 \omega_2 \varphi_2$, so that G is s.e. (since φ_1 and ψ_2 are non-null) contrary to assumption. Similarly, in case φ_2 and ψ_1 are non-null. Consequently G satisfies the antecedent condition of Lemma 1, and L is a finite state language.

(II) Suppose that G contains one nonterminal symbol S. Therefore either L is finite, hence (trivially) a finite state language, or infinite, in which case $S \rightarrow \varphi S \psi$, and L is a finite state language by (I).

(III) Suppose that Theorem 1 is true for all grammars containing less than n nonterminal symbols and that G contains n nonterminal symbols $A_1, \cdots, A_n$, where A_1 is the initial symbol S. By virtue of (I), we may

assume that for some particular j, there are no strings φ, ψ such that $A_j \Rightarrow \varphi A_1 \psi$.

Suppose that this $j \neq 1$. Let L′ be the language generated by G′ which differs from G only in that each rule $A_j \rightarrow \varphi$ of G is deleted and that A_j is replaced elsewhere in the rules by a terminal symbol b which is new. By inductive hypothesis, L′ is a finite state language. Furthermore, by inductive hypothesis, the set $K = \{x \mid A_j \Rightarrow x\}$ is a finite state language. Therefore, by Lemma 2, L is itself a finite state language.

Suppose that $j = 1$. Let $\varphi_1, \cdots, \varphi_r$ be the strings such that $A_1 \rightarrow \varphi_i$. For each $i \leqq r$, let $K_i = \{x \mid \varphi_i \Rightarrow x\}$. Suppose that $\varphi_i = \alpha_1 \cdots \alpha_m$. By inductive hypothesis, the set $L_j = \{x \mid \alpha_j \Rightarrow x\}$ is a finite state language. By Lemma 3, (1), K_i is therefore a finite state language. By Lemma 3, (2), L is therefore a finite state language.

This establishes the theorem.

RECEIVED: August 12, 1959.

REFERENCES

CHOMSKY, N. (1959). On certain formal properties of grammars. *Inform. and Control* **2**, 137–167.

KLEENE, S. C. (1956). Representation of events in nerve nets. *In* "Automata Studies" (C. E. Shannon and J. McCarthy, eds.), pp. 3–40. Princeton Univ. Press, Princeton, New Jersey.